How to Combine Visual Features with Tags to Improve Movie Recommendation Accuracy?

Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Farshad Bakhshandegan Moghaddam and Andrea Luigi Edoardo Caielli

Politecnico di Milano, Milan, Italy
{yashar.deldjoo,mehdi.elahi,paolo.cremonesi}@polimi.it
moghaddam@okit.de, andrea.caielli@mail.polimi.it
http://www.polimi.it

Abstract. Previous works have shown the effectiveness of using stylistic visual features, indicative of the movie style, in content-based movie recommendation. However, they have mainly focused on a particular recommendation scenario, *i.e.* , when a new movie is added to the catalogue and no information is available for that movie (*New Item* scenario). However, the stylistic visual features can be also used when other sources of information is available (*Existing Item* scenario).
In this work, we address the second scenario and propose a hybrid technique that exploits not only the typical content available for the movies (*e.g.*, tags), but also the stylistic visual content extracted form the movie files and fuse them by applying a fusion method called *Canonical Correlation Analysis (CCA)*. Our experiments on a large catalogue of 13K movies have shown very promising results which indicates a considerable improvement of the recommendation quality by using a proper fusion of the stylistic visual features with other type of features.

1 Introduction

Classical approaches to multimedia recommendation are of *unimodal* nature [35, 19, 7, 14]. Recommendations are typically generated based on two different types of item features (or attributes): metadata containing *High-Level* (or semantic) information and media entailing *Low-Level* (or stylistic) aspects.

The high-level features can be collected both from structured sources, such as databases, lexicons and ontologies, and from unstructured sources, such as reviews, news articles, item descriptions and social tags [6, 27, 29, 10, 28, 6, 11, 1, 2]. The low-level features, on the other hand, can be extracted directly from the media itself. For example, in music recommendation many acoustic features, *e.g.* rhythm and timbre, can be extracted and used to find perceptual similar tracks [3, 4].

In this paper, we extend our previous works on movie recommendation [15, 14, 13, 16, 12], where a set of low-level visual features were used to mainly address the new item cold start scenario [17, 34, 18]. In such a scenario, no information is available about the new coming movies (*e.g.* user-generated movies), and the

low-level visual features are used to recommend those new movies. While this is an effective way of solving the new item problem, visual features can be also used when the other sources of information is provided (*e.g.*, tags added by users) to the movies. Accordingly, a fusion method can be used in order to combine two types of features, *i.e.* low-level stylistic features defined in our previous works [15, 14] with user-generated tags into a joint representation in order to improve the quality of recommendation. Hence, we can formulate the research hypothesis as follows: *Combining the low-level visual features (extracted from movies) with tag features by a proper fusion method, can lead to more accurate recommendations, in comparison to recommendations based on these features when used in isolation.*

More particularly, we propose a *multimodal* fusion paradigm which is aimed to build a content model that exploits low-level correlation between visual-metadata modalities [1]. The method is based on *Canonical Correlation Analysis* (CCA) which belongs to a wider family of multimodal subspace learning methods known as *correlation matching* [23, 30]. Unlike very few available multimodal video recommender systems todate [36, 26] which treat the fusion problem as a basic linear modeling problem without studying the underlying spanned feature spaces, the proposed method learns the correlation between modalities and maximize the pairwise correlation.

The main contributions of this work are listed below:

- we propose a novel technique that combines a set of automatically extracted stylistic visual features with other source of information in order to improve the quality of recommendation.
- we employ a data fusion method called *Canonical Correlation Analysis (CCA)* [20, 21] that unlike traditional fusion methods, which do not exploit the relationship between two set of features coming from two different sources, achieves this by maximizing the pairwise correlation between two sets.
- we evaluate our proposed technique with a large dataset with more than 13K movies that has been thoroughly analyzed in order to extract the stylistic visual features[2].

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. Section 3 discusses the proposed method by presenting a description of the visual features and introducing a mathematical model for the recommendation problem and the proposed fusion method. Section 4 presents the evaluation methodology. Results and discussions are presented in Section 5. Finally, in Section 6 we present the conclusion and future work.

[1] Note that though textual in nature, we treat metadata as a separate modality which is added to a video by a community-user (tag) or an expert (genre). Refer to Table 1 for further illustration.

[2] the dataset is called *Mise-en-scene Dataset* and it is publicly available through the following link: `http://recsys.deib.polimi.it`

2 Related work

Up to the present, the exploitation of low-level features have been marginally explored in the community of recommender systems. This is while such features have been extensively studied in other fields such as computer vision and content-based video retrieval [32, 24]. Although for different objectives, these communities share with the community of recommender systems, the research problems of defining the "best" representation of video content and of classifying videos according to features of different nature. Hence they offer results and insights that are of interest also in the movie recommender systems context.

The works presented in [24, 5] provide comprehensive surveys on the relevant state of the art related to video content analysis and classification, and discuss a large body of low-level features (visual, auditory or textual) that can be considered for these purposes. In [32] Rasheed et al. proposes a practical movie genre classification scheme based on computable visual cues. [31] discusses a similar approach by considering also the audio features. Finally, in [37] Zhou et al. proposes a framework for automatic classification, using a temporally-structured features, based on the intermediate level of scene representation.

While the scenario of using the low-level features has been interesting for the goal of video retrieval, this paper addresses a different scenario, *i.e.*, when the the low-level features features are used in a recommender system to effectively generate relevant recommendations for users.

3 Method Descriptions

In this section, we present the proposed method.

3.1 Visual Features

Multimedia content in a video can be classified into three hierarchical levels:

- **Level 1 "High-level (semantic) features":** At this level, we have *semantic* features that deal with the concepts and events happening in a video. For example, the plot of the movie "The Good, the Bad and the Ugly", which revolves around three gunslingers competing to find a buried cache of gold during the American Civil War.
- **Level 2 "Mid-level (syntactic) features":** At the intermediate level we have *syntactic features* that deal with what objects exist in a video and their interactions. As an example, in the same movie there are Clint Eastwood, Lee Van Cleef, Eli Wallach, plus several horses and guns.
- **Level 3 "Low-level (stylistic) features":** At the lowest level we have *stylistic features* which define the Mise-en-Scene characteristics of the movie, *i.e.*, the design aspects that characterize aesthetic and style of a movie. As an example, in the same movie predominant colors are yellow and brown, and camera shots use extreme close-up on actors' eyes.

The examples above are presented for the visual modality as it forms the focus of our recent works [14, 15, 13]. In order to allow a fair comparison between different modalities, we present the hierarchical comparison of multimedia content across different modalities [7] in Table 1. Note that while the visual, aural and textual modalities are elements of the multimedia data itself, metadata is *added* to the movie after production.

Table 1: Hierarchical and modality-wise classification of multimedia features

Level	**Visual**	**Aural**	**Text**	**Metadata**
High	events, concepts	events, concepts	semantic similarity	summary, tag
Mid	objects/people, objects' interaction	objects/people, source	sentences, keywords	genre, tag, cast
Low	motion, color, shape, lighting	timbre, pitch spectral frequency	nouns, verbs, adjectives	genre, tag

Recommender systems in the movie domain typically use high-level or mid-level features such as genre or tag which appears in the form of metadata [25, 35, 19]. These feature usually cover a wide range in the hierarchical classification of content, for example tags most often contain words about events and incidents (high-level), people and places (mid-level) while visual features extracted and studied in our previous works (presented in Table 2) cover the low-level aspects. By properly combining the high-level metadata and low-level visual features we aim to maximize the informativeness of the joint feature representation.

3.2 Multimedia recommendation Problem

A multimedia document D (*e.g.* a video) can be represented with the quadruple

$$D = (d_V, d_A, d_T, d_M)$$

in which d_V, d_A, d_T are the *visual*, *aural* and *textual* documents constituting a multimedia document and d_M is the *metadata* added to the multimedia document by a human (*e.g.* tag, genre or year of production). In a similar manner, a user's profile U can be projected over each of the above modalities and be symbolically represented as

$$U = (u_V, u_A, u_T, u_M)$$

The multimedia components are represented as vectors in features spaces $\mathbb{R}^{|V|}$, $\mathbb{R}^{|A|}$, $\mathbb{R}^{|T|}$ and $\mathbb{R}^{|M|}$. For instance, $f_V = \{f_1, ..., f_{|V|}\}^T \in \mathbb{R}^{|V|}$ is the feature

Table 2: The list of low-level stylistic features representative of movie style presented in our previous works [14, 16]

Features	**Equation**	**Description**
Camera Motion	$\overline{L}_{sh} = \frac{n_f}{n_{sh}}$	Camera shot is used as the representative measure of camera movement. A shot is a single camera action. The Average shot length $\overline{L}_{sh}$ and number of shots n_{sh} are used as two distinctive features.
Color Variance	$\rho = \begin{pmatrix} \sigma_L^2 & \sigma_{Lu}^2 & \sigma_{Lv}^2 \\ \sigma_{Lu}^2 & \sigma_u^2 & \sigma_{uv}^2 \\ \sigma_{Lv}^2 & \sigma_{uv}^2 & \sigma_v^2 \end{pmatrix}$	For each keyframe in the Luv colorspace the covariance matrix ρ is computed where $\sigma_L, \sigma_u, \sigma_v, \sigma_{Lu}, \sigma_{Lv}, \sigma_{uv}$ are the standard deviation over three channels L, u, v and their mutual covariance. $\Sigma = det(\rho)$ is the measure for color variance. The mean and std of Σ over keyframes are used as the representative features of color variance.
Object Motion	$\triangledown I(x,t).v + I_t(x,t) = 0$	Object motion is calculated based on optical flow estimation. which provides a robust estimate of the object motion in video frames based on pixel velocities. The mean and std of of pixels motion is calculated on each frames and averaged across all video as the two representative features for object motion.
Lighting Key	$\xi = \mu.\sigma$	After transforming pixel to HSV colorspace the mean μ and std σ of the value component which corresponds to the brightness is computed. ξ which is the multiplication of two is computed and averaged across keyframes as the measure of average lighting key in a video.

vector representing the visual component. The relevancy between the target user profile U and the item profile D is of interest for recommenders and is denoted with $\mathcal{R}(U, D)$. In this work, we will focus our attention to the visual features defined in Table 2 and the rich metadata (tag).

Given a user profile U either directly provided by her (direct profile) or evaluated by the system (indirect profile) and a database of videos $\mathcal{D} = \{D_1, D_2, ..., D_{|D|}\}$, the task of video recommendation is to seek the video D_i that satisfies

$$D_i^* = \underset{D_i \in \mathcal{D}}{\arg\max}\ \mathcal{R}(U, D_i) \tag{1}$$

where $\mathcal{R}(U, D)$ can be computed as

$$\mathcal{R}(U, D_i) = R(F(u_V, u_M), F(d_V, d_M)) \tag{2}$$

where F is a function whose role is to combine different modalities into a joint representation. This function is known by *inter-modal fusion function* in multimedia information retrieval (MMIR). It belongs to the family of fusion methods

known as *early fusion* methods which integrates unimodal features before passing them to a recommender. The effectiveness of early fusion methods has been proven in couple of multimedia retrieval papers [30, 33].

3.3 Fusion Method

The fusion method aims to combine information obtained from two sources of features: (1)*LL Features*: stylistic visual features extracted by our system and (2)*HL features*: the tag features. We employ a novel fusion method known as *Canonical Correlation Analysis* (CCA) [30, 20, 21] for fusing two sources of features. CCA is popular method in multi-data processing and is mainly used to analyse the relationships between two sets of features originated from different sources of information.

Given two set of features $X \in R^{p\times n}$ and $Y \in R^{q\times n}$, where p and q are the dimension of features extracted from the n items, let $S_{xx} \in R^{p\times p}$ and $S_{yy} \in R^{q\times q}$ be the *between-set* and $S_{xy} \in R^{p\times q}$ be the *within-set* covariance matrix. Also let us define $S \in R^{(p+q)\times(p+q)}$ to be the *overall covariance matrix* - a complete matrix which contains information about association between pairs of features-represented as following

$$S = \begin{pmatrix} \text{cov}(x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{cov}(y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \tag{3}$$

then CCA aims to find a linear transformation $X^* = W_x^T.X$ and $Y^* = W_y^T.Y$ that maximizes the pair-wise correlation across two feature set as given by eq. 4. The latter will ensure the relationship between two set of features will follow a consistent pattern. This would lead to creation of discriminative and informative fused feature vector

$$\arg\max_{W_x,W_y} \text{corr}(X^*,Y^*) = \frac{\text{cov}(X^*,Y^*)}{\text{var}(X^*).\text{var}(Y^*)} \tag{4}$$

where $\text{cov}(X^*,Y^*) = W_x^T S_{xy} W_y$ and for variances we have that $\text{var}(X^*) = W_x^T S_{xx} W_x$ and $\text{var}(Y^*) = W_y^T S_{yy} W_y$. We adopt the maximization procedure described in [20, 21] and solving the eigenvalue equation

$$\begin{cases} S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \hat{W}_x = \Lambda^2 \hat{W}_x \\ S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} \hat{W}_y = \Lambda^2 \hat{W}_y \end{cases} \tag{5}$$

where $W_x, W_y \in R^{p\times d}$ are the eigenvectors and Λ^2 is the diagonal matrix of eigenvalues or squares of the *canonical correlations*. Finally, $d = rank(S_{xy}) \leq \min(n,p,q)$ is the number of non-zero eigenvalues in each equation. After calculating $X^*, Y^* \in R^{d\times n}$, feature-level fusion can be performed in two manners: (1)concatenation (2)summation of the transformed features:

$$Z^{ccat} = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x^T.X \\ W_y^T.Y \end{pmatrix} \tag{6}$$

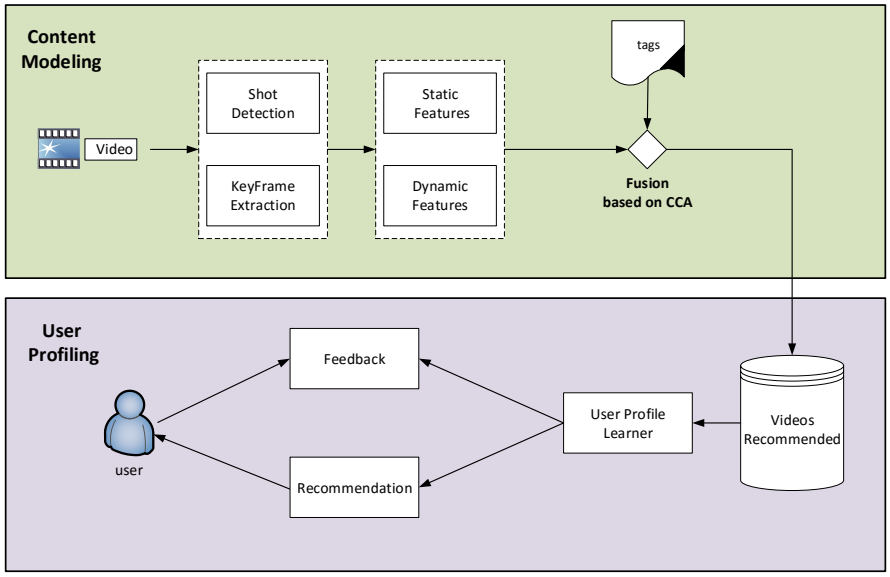

Fig. 1: Illustration of the proposed video recommender system based on stylistic low-level visual feature and user-generated tag using a fusion method based on CCA

and

$$Z^{sum} = X^* + Y^* = W_x^T.X + W_y^T.Y \tag{7}$$

Figure 1 illustrates the building block of the developed video recommender system. Color variance and lighting key are the extracted static features and camera and object motion are the dynamic features.

3.4 Recommendation algorithm

To generate recommendations, we adopted a classical "k-nearest neighbor" content-based algorithm. Given a set of users U and a catalogue of items I, a set of preference scores r_{ui} has been collected. Moreover, each item $i \in I$ is associated to its feature vector $\boldsymbol{f_i}$. For each couple of items i and j, a similarity score s_{ij} is computed using *cosine similarity* as follows

$$s_{ij} = \frac{\boldsymbol{f_i}^T \boldsymbol{f_j}}{\|\boldsymbol{f_i}\| \, \|\boldsymbol{f_j}\|} \tag{8}$$

For each item i the set of its nearest neighbors NN_i is built, $|NN_i| < K$. Then, for each user $u \in U$, the predicted preference score $\hat{r_{ui}}$ for an unseen item i is computed as follows

$$\hat{r_{ui}} = \frac{\sum_{j \in NN_i, r_{uj}>0} r_{uj} s_{ij}}{\sum_{j \in NN_i, r_{uj}>0} s_{ij}} \tag{9}$$

4 Evaluation Methodology

4.1 Dataset

We have used the latest version of Movielens dataset [22] which contains 22'884'377 ratings and 586'994 tags provided by 247'753 users to 34'208 movies (sparsity 99.72%). For each movie in Movielens dataset, the title has been automatically queried in YouTube to search for the trailer. If the trailer is available, it has been downloaded. We have found the trailers for 13'373 movies.

Low-level features have been automatically extracted from trailers. We have used trailers and not full videos in order to have a scalable recommender system. Previous works have shown that low-level features extracted from trailers of movies are equivalent to the low-level features extracted from full-length videos, both in terms of feature vectors and quality of recommendations [14].

We have used Latent Semantic Analysis (LSA) to better exploit the implicit structure in the association between tags and items. The technique consists in decomposing the tag-item matrix into a set of orthogonal factors whose linear combination approximates the original matrix [8].

4.2 Methodology

We have evaluated the Top-N recommendation quality by adopting a procedure similar to the one described in [9].

- We split the dataset into two random subsets. One of the subsets contains 80% of the ratings and it is used for training the system (train set) and the other one contains 20% of the rating and it is used for evaluation (test set).
- For each relevant item i rated by user u in test set, we form a list containing the item i and all the items not rated by the user u, which we assume to be irrelevant to her. Then, we formed a top-N recommendation list by picking the top N ranked items from the list. Being r the rank of i, we have a *hit* if $r < N$, otherwise we have a *miss*. Hence, if a user u has N_u relevant items, the precision and recall in its recommendation list of size N is computed.
- We measure the quality of the recommendation in terms of recall, precision and mean average precision (MAP) for different cutoff values $N = 5, 10, 20$.

5 Results

Table 3 represents the results that we have obtained from the conducted experiments. As it can be seen, both methods for fusion of LL visual features with

TagLSA features, outperform either of using the TagLSA and LL visual features, with respect to the all considered evaluation metrics.

In terms of recall, fusion of LL visual with TagLSA, based on concatenation of these features (ccat), obtains the score of 0.0115, 0.0166, and 0.0209 for recommendation at 5, 10, and 20, respectively. The alternative fusion method based on summation (sum), also scores better than the other baselines, *i.e.* , LL visual features and TagLSA, with the recall values of 0.0055, 0.0085, and 0.0112 for different recommendation sizes (cutoff values). These values are 0.0038, 0.0046, and 0.0053 for recommendation by using LL visual features and 0.0028, 0.0049, and 0.0068 for recommendation by using TagLSA. These scores indicate that recommendation based on fusion of LL visual features and TagLSA features is considerably better than recommendation based on these content features individually.

In terms of precision, again, the best results are obtained by fusion of LL visual features with TagLSA features based on concatenation with scores of 0.0140, 0.0115, and 0.0079 for recommendation at 5, 10, and 20, respectively. The alternative fusion method (sum) obtains precision scores of 0.0081, 0.0069, and 0.0048 which is better than the other two individual baselines. Indeed, LL visual features archives precision scores of 0.0051, 0.0037, and 0.0023, while the TagLSA achieves precision scores of 0.0045, 0.0041, and 0.0031. These results also indicates the superior quality of the recommendation based on fusion of LL visual features and TagLSA in comparison to recommendation each of these set of features.

Similar results have been obtained in terms of MAP metric. Hence, fusion method based on concatenation (ccat) performs the best in comparison to the other baselines, by obtaining the MAP scores of 0.0091, 0.0080, and 0.0076 for recommendation at 5, 10, and 20. The MAP scores are 0.0045, 0.0038, 0.0035 for fusion of features based on summation (sum), 0.0035, 0.0028, 0.0026 for LL visual features, and 0.0025, 0.0021, 0.0019 for TagLSA. Accordingly, the fusion of the LL visual features and TagLSA presents excellent performance in terms of MAP metric.

Overall, the results validates our hypothesis and shows that combining the Low-Level visual features (LL visual) extracted from movies with tag content, by adopting a proper fusion method, can lead to significant improvement on the quality of recommendations. This is an promising outcome and shows the great potential of exploiting LL visual features together with other sources of content information such as tags in generation of relevant personalised recommendation in multimedia domain.

6 Conclusion and Future Work

In this paper, we propose the fusion of visual features extracted from the movie files with other types of content (*i.e.*, tags), in order to improve the quality of the recommendation. In the previous works, the visual features are used mainly to solve cold start problem, *i.e.* , when a new movie is added to the catalogue

Table 3: Quality of recommendation w.r.t Recall, Precision and MAP when using low-level visual features and high-level metadata features in isolation compared with fused features using our proposed method based on Canonical Correlation Analysis.

Features	**Fusion Method**	**Recall**			**Precision**			**MAP**		
		@5	@10	@20	@5	@10	@20	@5	@10	@20
TagLSA	-	0.0028	0.0049	0.0068	0.0045	0.0041	0.0031	0.0025	0.0021	0.0019
LL	-	0.0038	0.0046	0.0053	0.0051	0.0037	0.0023	0.0035	0.0028	0.0026
LL + TagLSA	CCA-Sum	0.0055	0.0085	0.0112	0.0081	0.0069	0.0048	0.0045	0.0038	0.0035
LL + TagLSA	CCA-Ccat	**0.0115**	**0.0166**	**0.0209**	**0.0140**	**0.0115**	**0.0079**	**0.0091**	**0.0080**	**0.0076**

and no information is available for that movie. In this work, however, we use the stylistic visual features in combination with other sources of information. Hence, our research hypothesis is that a proper fusion of the visual features of movies may have led to a higher accuracy of movie recommendation, *w.r.t.* using these set of features individually.

Based on the experiments, we conducted on a large dataset of 13K movies, we successfully verified the hypothesis and shown that the recommendation accuracy is considerably improved when the the (low-level) visual features are combined with user-generated tags.

In future, we would consider fusion of additional sources of information, such as, audio features, in order to farther improve the quality of the content based recommendation system. Moreover, we will investigate the effect of different feature aggregation methods on the quality of the extracted information and on the quality of the generated recommendations.

References

1. C. C. Aggarwal. Content-based recommender systems. In *Recommender Systems*, pages 139–166. Springer, 2016.
2. C. C. Aggarwal. *Recommender Systems: The Textbook*. Springer, 2016.
3. D. Bogdanov and P. Herrera. How much metadata do we need in music recommendation? a subjective evaluation using preference sets. In *ISMIR*, pages 97–102, 2011.
4. D. Bogdanov, J. Serrà, N. Wack, P. Herrera, and X. Serra. Unifying low-level and high-level music similarity measures. *Multimedia, IEEE Transactions on*, 13(4):687–701, 2011.
5. D. Brezeale and D. J. Cook. Automatic video classification: A survey of the literature. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(3):416–430, 2008.
6. I. Cantador, M. Szomszor, H. Alani, M. Fernández, and P. Castells. Enriching ontological user profiles with tagging history for multi-domain recommendations. 2008.
7. O. Celma. Music recommendation. In *Music Recommendation and Discovery*, pages 43–85. Springer, 2010.

8. P. Cremonesi, F. Garzotto, S. Negro, A. V. Papadopoulos, and R. Turrin. Looking for good recommendations: A comparative evaluation of recommender systems. In *Human-Computer Interaction–INTERACT 2011*, pages 152–168. Springer, 2011.
9. P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010*, pages 39–46, 2010.
10. M. de Gemmis, P. Lops, C. Musto, F. Narducci, and G. Semeraro. Semantics-aware content-based recommender systems. In *Recommender Systems Handbook*, pages 119–159. Springer, 2015.
11. M. Degemmis, P. Lops, and G. Semeraro. A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction*, 17(3):217–255, 2007.
12. Y. Deldjoo, M. Elahi, and P. Cremonesi. Using visual features and latent factors for movie recommendation. In *Workshop on New Trends in Content-Based Recommender Systems (CBRecSys), in conjugation with ACM Recommender Systems conference (RecSys)*, 2016.
13. Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, and P. Piazzolla. Recommending movies based on mise-en-scene design. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1540–1547. ACM, 2016.
14. Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics*, pages 1–15, 2016.
15. Y. Deldjoo, M. Elahi, M. Quadrana, and P. Cremonesi. Toward building a content-based video recommendation system based on low-level features. In *E-Commerce and Web Technologies*. Springer, 2015.
16. Y. Deldjoo, M. Elahi, M. Quadrana, P. Cremonesi, and F. Garzotto. Toward effective movie recommendations based on mise-en-scène film styles. In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter*, pages 162–165. ACM, 2015.
17. M. Elahi, F. Ricci, and N. Rubens. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 20:29 – 50, 2016.
18. I. Fernández-Tobías, M. Braunhofer, M. Elahi, F. Ricci, and I. Cantador. Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction*, 26(2):221–255, 2016.
19. I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel. Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201. ACM, 2010.
20. M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi. Fully automatic face normalization and single sample face recognition in unconstrained environments. *Expert Systems with Applications*, 47:23–34, 2016.
21. D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
22. F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2015.
23. H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

24. W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819, 2011.
25. X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *Proceedings of the 17th international conference on World Wide Web*, pages 675–684. ACM, 2008.
26. T. Mei, B. Yang, X.-S. Hua, and S. Li. Contextual video recommendation by multimodal relevance and user feedback. *ACM Transactions on Information Systems (TOIS)*, 29(2):10, 2011.
27. R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204. ACM, 2000.
28. M. Nasery, M. Braunhofer, and F. Ricci. Recommendations with optimal combination of feature-based and item-based preferences. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 269–273. ACM, 2016.
29. M. Nasery, M. Elahi, and P. Cremonesi. Polimovie: a feature-based dataset for recommender systems. In *ACM RecSys Workshop on Crowdsourcing and Human Computation for Recommender Systems (CrawdRec)*, volume 3, pages 25–30. ACM, 2015.
30. J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):521–535, 2014.
31. Z. Rasheed and M. Shah. Video categorization using semantics and semiotics. In *Video mining*, pages 185–217. Springer, 2003.
32. Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(1):52–64, 2005.
33. N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.
34. N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan. Active learning in recommender systems. In *Recommender Systems Handbook*, pages 809–846. Springer, 2015.
35. M. Szomszor, C. Cattuto, H. Alani, K. OHara, A. Baldassarri, V. Loreto, and V. D. Servedio. Folksonomies, the semantic web, and movie recommendation. 2007.
36. B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 73–80. ACM, 2007.
37. H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg. Movie genre classification via scene categorization. In *Proceedings of the international conference on Multimedia*, pages 747–750. ACM, 2010.